quantitative measures such as the target registration error can be used during commissioning, such measures are not fully spatial and too user intensive in clinical practice. Therefore, we propose a fully automatic and quantitative approach to DIR quality assessment including multiple measures of numerical robustness and biological plausibility.

Material and Methods: Ten head and neck cancer patients who received weekly repeat CT (rCT) scans were included. Per patient, the first rCT was deformable registered (using B-spline DIR algorithm) to the planning CT. The ground-truth deformation error of this registration was derived using the scale invariant feature transform (SIFT), which automatically extracts and matches stable and prominent points between two images. Moreover, complementary quantitative and spatial measures of registration quality were calculated. Numerical robustness was derived from the inverse consistency error (ICE), transitivity error (TE), and distance discordance metric (DDM). For the TE calculations a third CT was used. The DDM was calculated using five CT sets per patient. Biological plausibility was based on the deformation vector field between the planning CT and rCT. Relative deformation threshold values were set based on physical tissue characteristics: 5% for bone and 50% for soft tissues. All measures were evaluated in bone and soft tissue structures and compared against the ground-truth deformation error.

Results: On average, SIFT detected 133 matching points scattered throughout the planning CT, with a mean (max) registration error of 1.6 (8.3) mm. Our combined and fully spatial DIR evaluation approach, including the ICE, TE and DDM, resulted in a mean (max) error of respectively 0.6 (2.0), 0.7 (2.7), and 0.6 (2.7) mm within the external body contour, averaged over all patients. The largest errors were detected in homogeneous regions and near air cavities. Furthermore, 87% of the bone and 2% of the soft tissue voxels were classified as unrealistic deformations. Figure 1 shows the planning CT, DDM, tissue deformation, and error volume histograms of the ICE, TE, and DDM of the body contour of one patient.

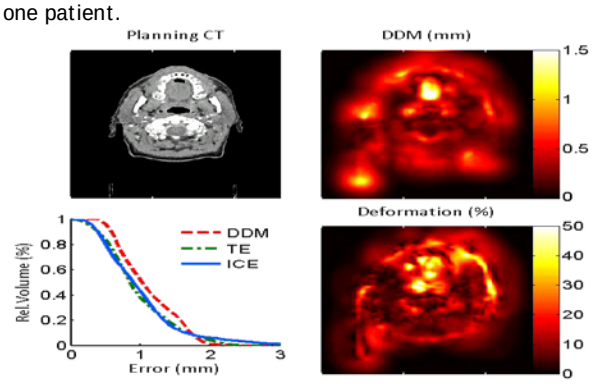


Figure 1. Axial cross section of a planning CT, distance discordance metric (DDM) and tissue deformation of a representative patient. The error volume histograms show the DDM, transitivity error (TE), and inverse consistency error (ICE) of the body contour.

Conclusion: The combination of multiple automatic DIR quality measures highlighted areas of concern within the registration. While current methods on DIR evaluation, such as visual inspection and target registration error are time-consuming, local, and qualitative, this approach provided an automated, fully spatial and quantitative tool for clinical assessment of patient-specific DIR even in image regions with limited contrast.

OC-0068

Can atlas-based auto-contouring ever be perfect?

B.W.K. Schipaanboord[1], J. Van Soest[2], D. Boukerroui[1], T. Lustberg[2], W. Van Elmpt[2], T. Kadir[1], A. Dekker[2], M.J. Gooding[1]

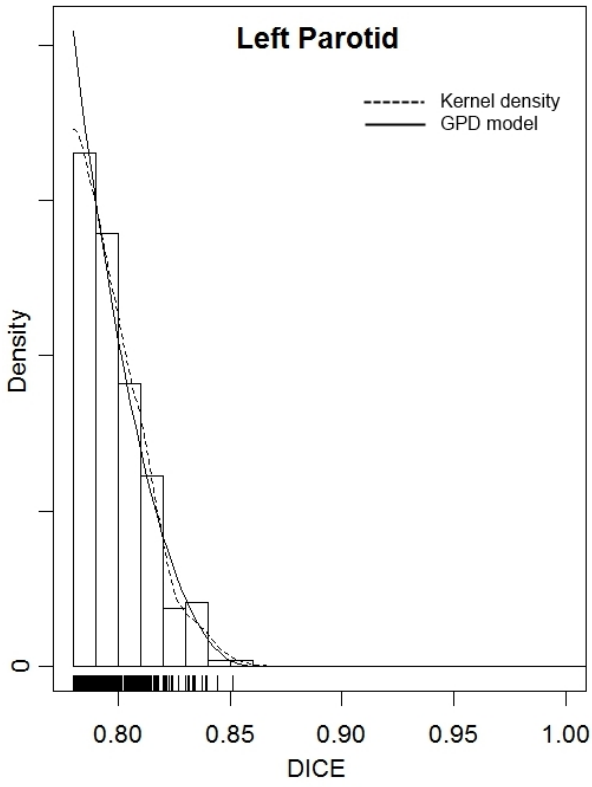
[1]*Medical Ltd, Science and Medical Technology, Oxford, United Kingdom*

[2]*Maastricht University Medical Centre, Department of Radiation Oncology MAASTRO- GROW School for Oncology and Developmental Biology, Maastricht, The Netherlands*

Purpose or Objective: Various approaches have been proposed to select the most similar atlases to a patient for atlas-based auto-contouring. While it is known that increasing the size of an atlas database improves the results of auto-contouring for a small number of atlases, such selection assumes the hypothesis that increasing the atlas pool size always increases the chance of finding a good match. The objective of this study is to test this hypothesis, and answer the question; "Given a large enough database of atlases, can single atlas-based auto-contouring ever be perfect?".

Material and Methods: 35 test cases were randomly selected from a dataset of 316 clinically contoured head and neck cases, and were auto-contoured treating each of the remaining cases as potential atlases to be used. Thus, results of contouring were available for approximately 11000 atlas-patient pairs. Dice Similarity Coefficient (DSC), Hausdorff distance (HD), Average Distance (AD) and Root Mean Square Distance (RMSD) were computed between the auto-contours and the clinical contours for each structure and atlas-patient pair. In order to estimate achievable performance under the assumptions of an infinite size atlas database and "perfect" atlas selection, the Extreme Value Theory statistical technique Points over Threshold, used in other domains to perform tasks such as estimating the magnitude of one-in-a-hundred-years flooding, was used to model the distribution of the best scores. Analysis was performed for the ten most commonly contoured structures within the database, with a minimum of 6800 atlas-patient pairs per structure being considered.

Results: The figure shows the distribution of observed extreme values for the left parotid DICE scores, together with the model fit.



For all measures and structures, the model fit indicated a limit on the performance in the extreme. While this is expected since all measures have a limit at perfection, the performance limit in the extreme fell short of a perfect result. Variation was observed between structures, with well-defined structures performing better than more complex ones. This may indicate that the limit on performance reflects the inter-observer variation in delineation. The table shows the best observed score for the experiments performed, together with the expected achievable result predicted by the model assuming an atlas database of 5000 atlases.

Structure	DICE		HD (mm)		AD (mm)		RMSD (mm)	
	Best	Exp	Best	Exp	Best	Exp	Best	Exp
Brain	0.987	0.997	5.98	5.98	0.54	0.52	0.86	0.83
Brainstem	0.886	0.887	4.58	4.50	1.15	1.09	1.43	1.34
Cochlea L	0.909	0.935	0.48	0.47	0.19	0.17	0.22	0.21
Cochlea R	0.909	0.945	0.52	0.47	0.19	0.17	0.23	0.20
Oral Cavity	0.894	0.894	8.07	8.13	1.79	1.75	2.34	2.31
Parotid L	0.852	0.854	6.12	5.68	1.32	1.31	1.70	1.70
Parotid R	0.849	0.853	6.10	6.09	1.41	1.34	1.89	1.83
Spinal cord	0.881	0.881	2.92	3.05	0.65	0.65	0.86	0.82
Submandibular gland L	0.840	0.845	4.88	4.79	1.10	1.05	1.40	1.35
Submandibular gland R	0.841	0.842	4.16	4.04	1.02	1.03	1.34	1.34

Conclusion: Increasing the size of an atlas database within achievable ranges would be insufficient on its own for consistently perfect single atlas auto-contouring, even in the presence of a "perfect" atlas selection method. Thus, improvements in the underlying methods for pre- and post-processing, such as deformable registration or multi-atlas fusion, are necessary to improve the results of atlas-based auto-contouring. Additionally, consistent delineation within an atlas database is required to minimise the effect of inter-observer error on the achievable performance.

OC-0069

Using texture analysis to detect prostate cancer for automated outlining and adaptive radiotherapy

D. Welsh[1], D. Montgomery[1], D.B. McLaren[2], W.H. Nailon[1]
[1]*Western General Hospital, Oncology Physics, Edinburgh, United Kingdom*
[2]*Western General Hospital, Clinical Oncology, Edinburgh, United Kingdom*

Purpose or Objective: In radiotherapy, the prostate is one of few anatomical sites where the whole organ is targeted, even in cases of localised cancer. Improvements in outcomes may be achieved by escalating the dose to the dominant intraprostatic lesion (DIL), and thereby reducing the dose to the remainder of the gland. However, reliably identifying the DIL requires considerable clinical experience and is extremely time consuming. Automated outlining would alleviate this problem, and is also desirable for online adaptive radiotherapy. This work investigated the feasibility of automatically detecting the DIL on T2-weighted MR images using image texture analysis methods.

Material and Methods: On the diagnostic T2-weighted MR images from 14 prostate cancer patients previously treated with radiotherapy, the prostate and DIL volumes were defined by a clinician. Two separate projects were carried out using the same data, looking at 2D and 3D texture analysis, respectively. In both cases, a range of texture features were calculated on a sub-volume basis and a machine learning classification scheme was trained to classify individual pixels surrounded by each sub-volume as either healthy prostate or DIL, based on the calculated features, with the clinician defined contours as the ground truth. The classifier was tested on each patient case in turn, with the remaining 13 patients used as the training data in a leave-one out schema. Classification results were assessed in terms of receiver operator characteristic (ROC) and confusion statistics.

		Sensitivity	Specificity	aucROC
2D	Mean	0.43 ± 0.34	0.91 ± 0.09	0.82 ± 0.13
	Min	0.00	0.68	0.54
	Max	0.95	0.99	0.95
3D	Mean	0.43 ± 0.27	0.74 ± 0.12	0.60 ± 0.16
	Min	0.02	0.56	0.26
	Max	0.93	0.94	0.80

Table 1: Summary of classifier performance across the 14 patient cases.

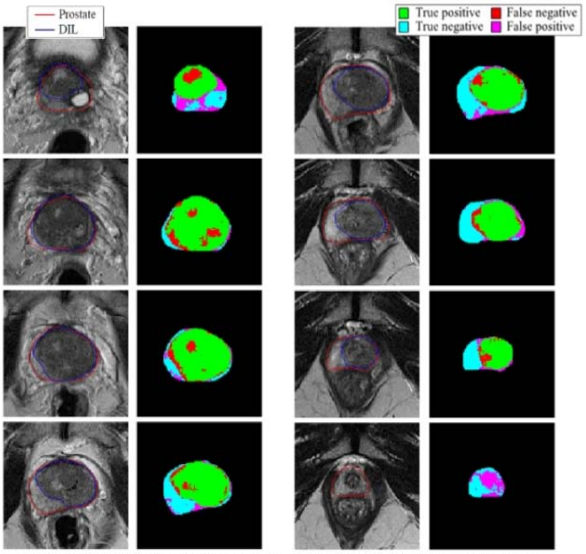


Figure 1: Illustration of classifier results for an example case, with colour map of results shown alongside corresponding MR image slice.

Results: Over the 14 patients, the best performing 2D analysis resulted in a mean area under the ROC curve (aucROC) of 0.82 ± 0.13, whilst the 3D analysis gave an aucROC of 0.60 ± 0.16. A summary of the results is shown in Table 1 and Figure 1 shows a visualisation of the (2D) classification results for an example case. There is wide variation in classifier performance from case to case - performance tended to be poorer on patients with small DILs, giving a low sensitivity but high specificity. The mean value of sensitivity is heavily affected by these low scoring cases. It is expected that the results could be improved with a larger training dataset and morphological post-processing of the detected DIL region.

Conclusion: This work shows that, in principle, texture analysis can be used to identify focal lesions on MR images, facilitating automated delineation for adaptive radiotherapy. 3D analysis does not necessarily lead to improved performance over 2D, although further optimisation of both methods may be possible.

OC-0070

Do radiomics features excel human eye in identifying an irradiated tumor? Rat tumor to patient HNSCC

K. Panth[1], S. Carvalho[1], A. Yaromina[1], R. T.H. Leijenaar[1], S. J. Van Hoof[1], N. G. Lieuwes[1], B. Rianne[1], M. Granzier-Peeters[1], F. Hoebers[1], D. Eekers[1], M. Berbee[1], L. Dubois[1], P. Lambin[1]
[1]*MAASTRO clinic, Radiation Oncology, Maastricht, The Netherlands*

Purpose or Objective: Radiomics hypothesizes that imaging features reflect the underlying gene expression patterns and intratumoral heterogeneities. In this study, we hypothesized that radiation treatment (RT) affects image features and that